CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Validating and Calibrating a Destination Estimation Algorithm for Public Transport Smart Card Fare Collection Systems

Li He
Neema Nassir
Martin Trépanier
Mark Hickman

October 2015

CIRRELT-2015-52

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

UNIVERSITÉ LAVAL · McGill · Concordia · ÉTS · UQÀM · HEC MONTRÉAL · POLYTECHNIQUE MONTRÉAL · Université de Montréal

Validating and Calibrating a Destination Estimation Algorithms for Public Transport Smart Card Fare Collection Systems

Li He[1], Neema Nassir[2], Martin Trépanier[1,3,*], Mark Hickman[2]

[1] Department of Mathematics and Industrial Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

[2] University of Queensland, School of Civil Engineering, room 555, AEB (Bldg 49), St Lucia, Queensland 4072, Australia

[3] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

Abstract. Data from smart card fare collection systems has proven to be very useful to public transport planners. These systems provide a continuous flow of data on transactions made on networks; hence it helps to better understand customer (card) travel behavior, and the data can also be used to characterize and model general ridership, customer loyalty, and network performance indicators. But many systems only record the entrance ("tap-in") transaction in the system. There is a need to estimate the exit ("tap-out") location to have origin-destination trip information. In this paper, we use tap-in/tap-out smart card data from Brisbane, Australia, to calibrate and validate a trip destination estimation algorithm developed for Canadian data. Results show that the algorithm has an accuracy of 79% within an acceptable distance of 400 m. The proposed calibration method helped to solve 1.4% more destinations.

Keywords: Public transport, smart card fare collection system, trip modelling.

Acknowledgements. The authors wish to acknowledge the supporters of this study, which are the Société de Transport de l'Outaouais, the Natural Science and Engineering Research Council of Canada (project RDCPJ 446107 12), Thalès Research and Technologies, and the Queensland Department of Transport and Main Roads.

Results and views expressed in this publication are the sole responsibility of the authors and do not necessarily reflect those of CIRRELT.

Les résultats et opinions contenus dans cette publication ne reflètent pas nécessairement la position du CIRRELT et n'engagent pas sa responsabilité.

---

* Corresponding author: Martin.Trepanier@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2015

© He, Nassir, Trépanier, Hickman and CIRRELT, 2015

1 Introduction

Nowadays, smart card data are very useful for public transport planning. Smart card automated fare collection systems provide a continuous flow of data on transactions made on networks; hence it helps to better understand customer (card) travel behavior, and the data can also be used to characterize and model general ridership, customer loyalty, and network performance indicators.

However, many smart card systems only record the boarding transactions ("tap-in") and not the alighting transactions ("tap-out"). Origins and destinations are essential to public transport planners because most of their models work with origin-destination matrices that describe the movements of demand. In "tap-in" only systems, algorithms are needed to estimate the destination location. However, at this time, there have been no studies that provide direct validation of the results of these estimates: the resulting inferred destinations often cannot be independently confirmed.

In this paper, we propose to calibrate and validate a destination estimation algorithm that was developed for Gatineau, Canada with the help of "tap-in/tap-out" data from the region of Brisbane, Australia. This exercise helps to validate the simplest assumptions of the algorithm (based on the sequence of stops) and calibrates its parameters. It also helps to improve the development of the algorithms that are related to unlinked trips that are processed using historical data.

The paper first presents some background on the usage of smart card data in public transport planning and also presents works related to origin-destination estimation from smart card data. Then, the "methodology" section details the case study and the algorithm that was used, plus the parameters to be calibrated. The "experiments" section reports on the main results of the validation. It first emphasizes on the validation of the threshold distance, i.e. the distance for which we "accept" the destination. Then, it presents the results according to the algorithm and the temporal attributes of the trips.

2 Background

2.1 Smart card data in public transit

In public transport, smart card automated fare collection systems (SCAFCS) are mainly used for revenue collection. The main workflow of these systems is simple:

1) a public transport user acquires a smart card and buys a fare or puts money on the card;
2) when the user enters a vehicle, he taps his card on a reader that validates the fare, and the time of this "tap-in" transaction is recorded (date, time, location, bus number, router number, direction, etc.);
3) in some systems, the user must validate when he exits the vehicle, where this "tap-out" transaction is recorded; and,
4) the vehicle or station data (for rail service) is downloaded regularly to a central server for revenue processing.

In addition to revenue management, there are many works that have proven that more can be done with smart card data in public transit planning at strategic, tactical and operational levels (Pelletier et al. 2011). Due to the huge quantity of data, earlier works have tried to classify passengers using data mining techniques (Morency et al. 2007). Devillaine et al. (2012) have shown a method to identify daily activities from smart card data, comparing networks from Santiago, Chile and Gatineau, Canada. Spurr et al. (2014) have proposed a method to correct household surveys using smart card data in the Montreal region. Because they represent a high density of transactions made in buses, smart card data can also be used to calculate public transport network (supply-side) performance indicators, as demonstrated by Trepanier et al. (2009).

Data from SCAFCS are also very useful for modeling. Shimamoto (2014) has used such data to calibrate a hyperpath choice model for public transport users. Smart card data are also used to model the influence of weather on demand (Trépanier et al. 2012) and to estimate the retention rate as a proxy to measure the loyalty of public transport users with respect to the service (Trépanier and Morency 2010). Data can be used to model very precise elements, like the waiting and walking times of individuals in a subway system (Lee and Ali 2014).

2.2 Destination estimation and OD matrices

As a matter of fact, many of the studies need OD matrices derived from smart card data, or at least part of these data. That is, methods are needed and have been developed to estimate the alighting transaction locations in the case of "tap-in" only networks.
A method based on the sequence of boarding stops has been proposed by Trépanier et al. (2007) for the smart card data of the Société de Transport de l'Outaouais (STO, Gatineau, Canada). This method assumes that users will alight at the stop which is the nearest to his subsequent boarding stop on the same day. For the last trip of the day, the boarding location of the trip of the next day is used to find the alighting point. Munizaga and Palma (2012) have proposed an improvement to the method to distinguish "real" transfers from hidden activities; this helps to better identify trips during a journey. Gordon et al. (2012) have proposed spatio-temporal criteria to differentiate transfers from activities between transactions. Nassir et al. (2015) also proposed an algorithm for distinguishing the transfers from short activities by comparing the recorded travel time between the origin-destination pairs with the fastest transit options. He and Trepanier (2015) improved the estimation of alighting stops of unlinked trips using a kernel density estimation of probabilities of alighting. This method is presented hereafter in the "Methodology" section.
Smart card data can also be completed by fusing with other sources of data. Nassir et al. (2011) have produced origin-destination matrices on public transport network using General Transit Feed Specification (GTFS) schedule data, automated passenger counting (APC) data, and automated vehicle location (AVL) data.

3 Methodology

3.1 Data source

The calibration and validation data for this study are taken from a smart card data set from Brisbane, Australia. These data represent both tap-on and tap-off data from the Go Card, a smart card which is used for approximately 85% of the public transport journeys made in Brisbane on any given day. The dataset used in this study is a subset of 40,431 trips made during the month of March 2013 (further detail about the dataset is masked to ensure privacy). These trips were made by a random set of card users.

3.2 Destination estimation algorithm

The destination estimation algorithm that was used in this study has a mathematical formulation that can be found in He and Trépanier (2015). The algorithm is divided in two parts: 1) estimation of alighting stops based on stop sequences (i.e. trip chaining assumptions), and 2) alighting stop estimation for unlinked trips based on longitudinal analysis of the smart card data.
For the first part, the sequence of boarding stops during a journey is used to find the most probable alighting stop. In Figure 1, the estimated alighting stop of the first route is located at the stop which is the nearest to the next boarding stop (on route 2). It is the same case for alighting stop 3, which is the nearest to the first boarding of the day. However, we cannot determine the second alighting stop because the

distance d between all stops of route 2 and the boarding of route 3 is over 2000 meters, which is the default "tolerance distance" of the algorithm. This distance will be calibrated in our case study.

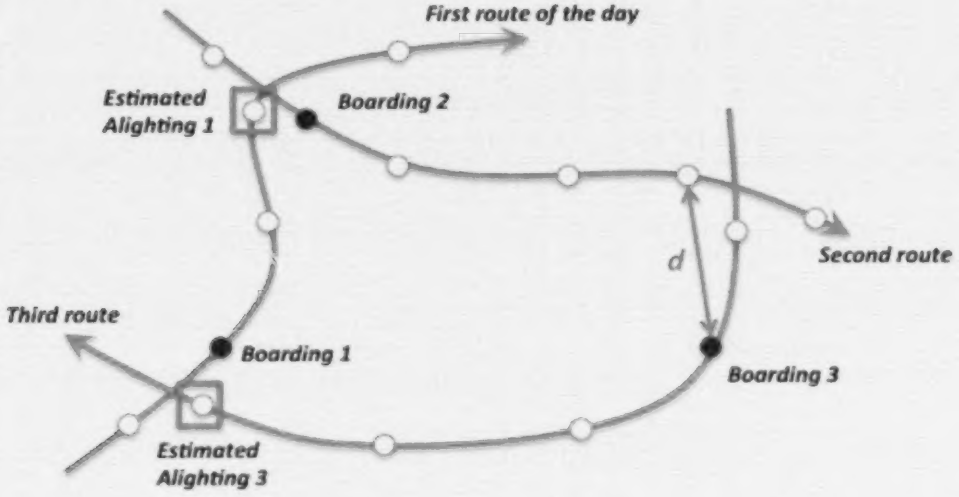

Figure 1. Part 1 of the destination estimation algorithm

3.3 Destination estimation algorithm

The second trip of the example journey in Figure 1 is said to be "unlinked", because we cannot find the alighting based on the sequence of transactions. The same applies to journeys with only one trip. The estimation of the alighting stops of these unlinked trips is done in the second part of the algorithm, which uses the historical records of transactions of the individual card. For example, for a given unlinked trip, we investigate the set of all other transactions made by the same card, at the same boarding stop, and in the same route direction. We utilize the already estimated alighting stops for this set of transactions in order to infer the most probable alighting stop for an unlinked trip. The logic behind this method is that passengers are assumed to have repetitive patterns of travel and activity locations at regular times, and this pattern can be identified and exploited for estimating the destination stops of the unlinked trips. In this algorithm, for each stop in this historical set, we use a kernel density estimator to estimate the probability of the expected arrival time and distance between the boarding stop and possible alighting stop. Then, we multiply the time and distance probabilities for each possible alighting location. The location with the highest probability is inferred as the alighting stop.

This algorithm uses basic parameters that have been fixed by He and Trépanier (2015) because no "tap-in/tap-out" data was available for validation. These parameters include: the maximum distance d for the first part of the algorithm, the threshold distance and time for the second part, and the hypothesis that time and space are put on the same level of importance (probabilities are multiplied). The availability of "tap-in/tap-out" data makes it possible to test different combinations of parameters, thus helping to calibrate the algorithm and validate its results.

The algorithm has been implemented in Python, thus ensuring calculation automatically among a large amount of data. In this case, our input files are transaction data (including trip ID, card ID, date, time, bus line, direction, boarding stop), "ssli" data (including bus line, direction, stop, distance from first stop to given stop), and "stop data" (including stop, coordinates). With these input files and the python program, we obtain an output file that contains estimated destinations.

At the end, we use the following codes to analyze the results of the algorithm:

- Code 11: Destination is found in the trip sequence phase.
- Code 12: Destination is found in the last trip of day phase (also called return home).
- Code 13: Destination is found in the first trip of next day phase.
- Code 21: Destination is found with the kernel density method, where the unlinked trip has several potential destinations.
- Code 22: Destination is found with the kernel density method, but the unlinked trip has only one potential destination.

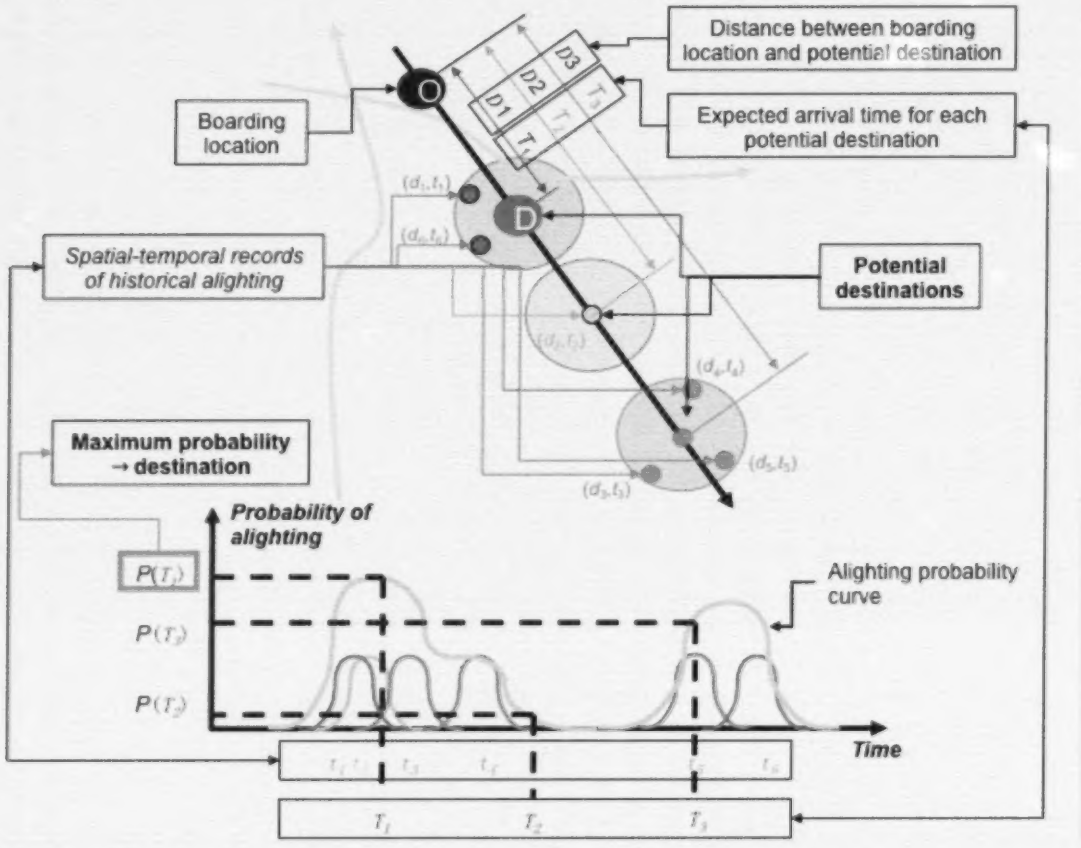

Figure 2. An illustration of the method to determine the most probable destination base on history of the card (from He and Trépanier 2015)

3.4 Distance threshold for accuracy

Not all the estimated destinations may match exactly the observed alighting stops. In a given bus route, stops may be relatively close to each other. For example, the estimated destination stop may be 200 meters from the observed destination stop. In many planning tools, such a threshold is rather acceptable, especially if OD matrices are based on zones that can be much larger in size. The idea here is to measure the accuracy of the algorithm within an acceptable distance. Our hypothesis is that the accuracy will increase rapidly with the increase of the acceptable distance, and then it may stabilize. Our objective is to find a suitable threshold for this parameter.

4 Experiments

One may first look at the overall accuracy of the destination estimation algorithm as it was developed by Li and Trépanier (2015), using tolerance distances of 2000 m for phases 11, 12 and 13, and 1000 m for phases 21 and 22. This is done to apply the base case from Gatineau to the Brisbane data. These results are analyzed accordingly to the phase of the algorithm that was used to find the destination. Results are also examined by day of the week and hour of the day.

4.1 Accuracy

We define the accuracy as the ability of the algorithm to find the destination within an acceptable distance from the true destination. The accuracy is 65.8% when the acceptable distance is 0 m. Logically, as we increase the acceptable distance, the accuracy will also increase. Figure 3 presents the relation between the acceptable distance and the accuracy of the proposed algorithm when applied to the Brisbane data set.

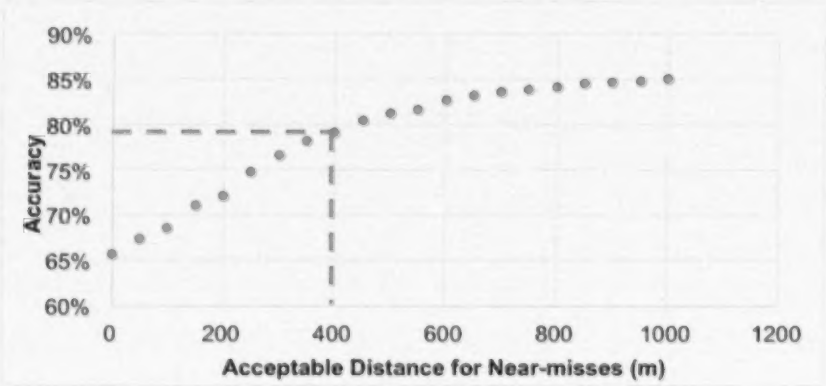

Figure 3 Calibration for acceptable distance

The curve shows that the accuracy increases almost linearly for a distance between 0 to 400 meters, and then begins to have more marginal improvement. There seems to be stability near 85% accuracy over 1000 meters, but theoretically 100% accuracy will be attained at a very large distance, because the set of destination is always chosen within a finite set of stops that stem from the boarding stop on a given route.
In the following result presentation, we use the acceptable distances of 0 m (perfect match) and 400 m (threshold distance) for the analyses.

4.2 Estimation phases

Figure 4 shows the estimation accuracy of different phases of the algorithm for the acceptable distances of 0 m and 400 m. The list of codes is presented in the methodology section.

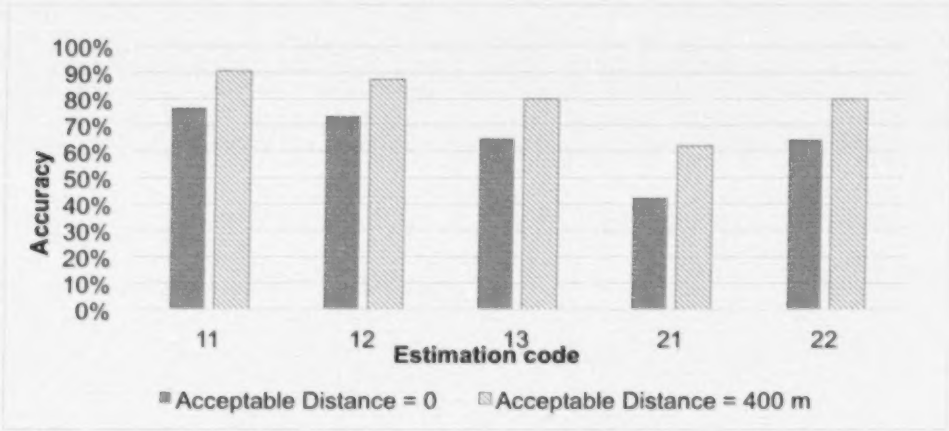

Figure 4 Estimation accuracy of each code for different acceptable distances

As expected, code 11 has the highest accuracy because the estimation is based on the trip sequence within a day. The method to estimate unlinked trips has a lower accuracy when there are several potential alighting stops (code 21): the accuracy is just above 40%. However, for the case where there is only one potential alighting stop in the historical transactions (code 22), the accuracy is similar to the first three phases.

With an acceptable distance of 400 m, the accuracy of each phase is above 60%, especially for code 11, where the accuracy is over 90%. It is noteworthy that for code 21, the accuracy increase is substantial for 400 meters distance. This means there are many observed destinations that are not so far away from the estimated destinations. It is interesting then to find a way to improve the method for code 21 to improve the estimation method. It is possible to improve the phase 21 because estimated destinations are already not far from the real destinations.

4.3 Temporal analysis

In this section, the accuracy is investigated by time of day and day of week, because the algorithm is based on the trip sequences (or temporal kernel estimates) which may vary by time and day of travel. Therefore, we propose two temporal analyses: (1) by time of day, and (2) by day of week.

In figure 5, apart from the 5:00 am peak that may not be significant, the highest accuracy of the perfect match (acceptable distance = 0 m) occurs in the evening peak (around 16:00). This would possibly mean that the behavior of returning home is more regular (the final stop would not vary). For the morning peak, the difference between the results of the two acceptable distances is high; this could be because passengers have a tendency to choose different stops to get off when they go to work or school. Since smart card data does not include any socio-demographic attributes attached to each transaction, we cannot further analyze the trip data by trip purpose or by population segment.

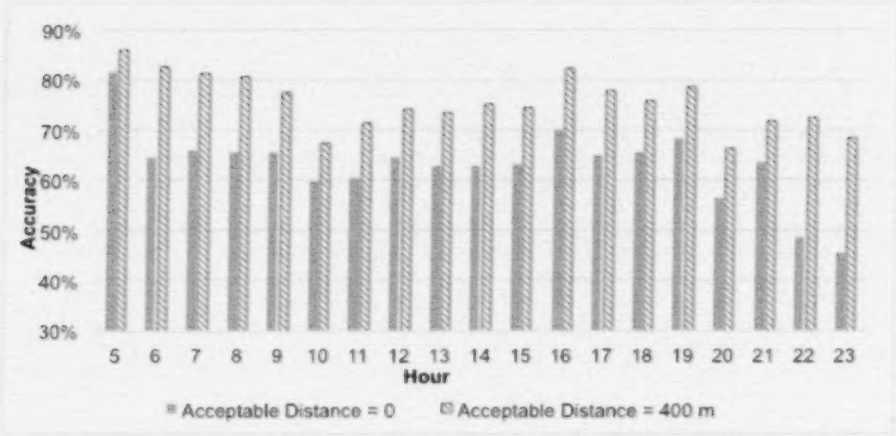

Figure 5 Estimation accuracy by time of day for different acceptable distances

Figure 6 shows the variation of the accuracy according to the day of the week. The accuracy in the weekdays is relatively higher than in the weekend. In the weekdays, the trips are more regular, as they follow the pattern of going to work and returning home. Hence, the estimation of the destination in the weekday is more accurate than in the weekend. However, it is interesting to note that the accuracy of estimation for Sunday transactions is higher than Saturday, maybe because people are more likely to return home on Sundays to prepare for their first trip on Mondays. Therefore, phase 13, which is based on trip chaining with the next day's information, can be more efficient for the transactions made on Sundays as compared to Saturdays.

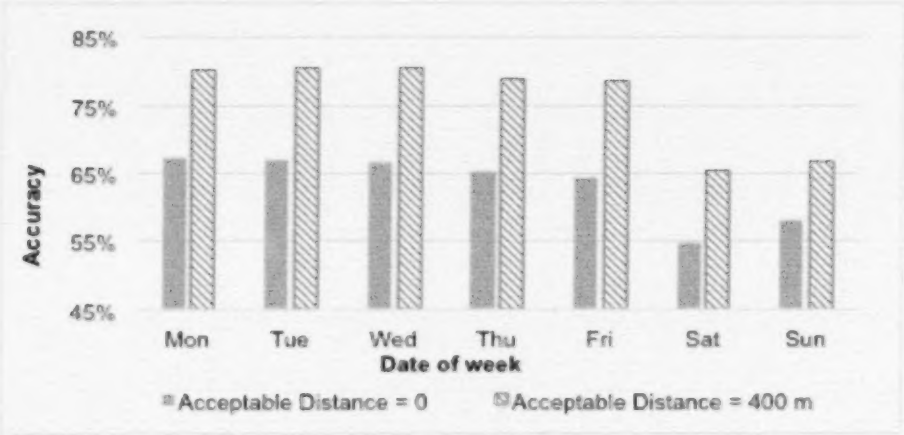

Figure 6 Estimation accuracy by day of week for different acceptable distances

5 Tolerance distance calibration

In the classical model, the tolerance distance defines the area in which the stops will be accepted to link to the next route. However, there was no method to calibrate this parameter. In this paper, we use Brisbane smart card data to try different distance values, aiming to find the best tolerance distance. We use two values of tolerance distance: a 2000 m threshold for phases 11, 12 and 13, and a 1000 m threshold for

phases 21 and 22. Because the estimation algorithm runs step by step, we first try different tolerance distances for phases 11, 12 and 13 and keep the distance values for phases 21 and 22 unchanged. After finding the best parameter value for phases 11, 12 and 13, we then try to calibrate for phases 21 and 22 by using the same method.

5.1 Tolerance distance for phases 11, 12 and 13

Figure 7 shows the accuracy of estimation when the tolerance distance ranges from 500 m to 2500 m. We can find that for all 3 phases, the accuracy decreases with the increase of tolerance distance. It seems that 500 m is the best choice. However, as shown in Figure 8, the quantity of trips estimated will be significantly decreased at 500 m. There is obviously a trade-off between the quality of estimation and the quantity of output. As a result, the optimal distance threshold that could increase the accuracy of estimation and would not significantly compromise the number of estimated stops could be1000 m, because the number of trips estimated at this distance is almost similar to the baseline case (2000 m). At this distance, the accuracy remains acceptable.

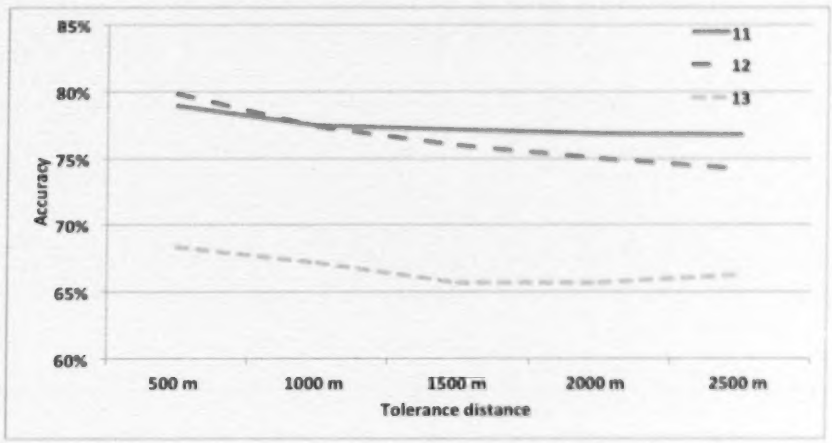

Figure 7 Calibration for tolerance distance of phases 11,12 and 13 (1) – estimated accuracy

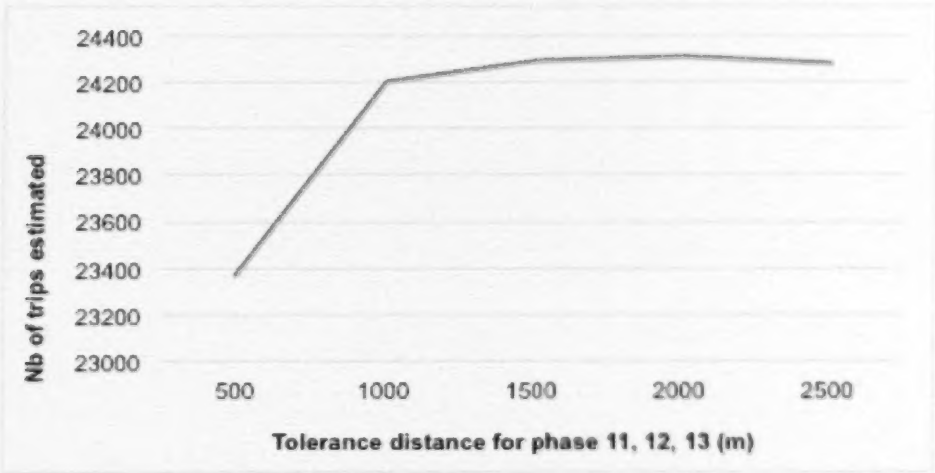

Figure 8 Calibration for tolerance distance of phases 11, 12 and 13 (2) – total trips estimated

5.2 Tolerance distance for phases 21, 22

For calibrating the threshold distance for phases 21 and 22, we keep constant the parameter value (1000m) that was found for the trip chaining phases 11, 12 and 13. We try four values for threshold distance of phases 21 and 22: 250 m, 500 m, 1000 m (baseline value) and 1500 m. Figure 9 shows the accuracy of estimation using these values. Obviously, the accuracy increases with the decrease of the tolerance distance. But is the number of estimated trips decreasing with the decrease of tolerance distance? We can find in Figure 10 that the number of trips estimated is not significantly decreased when the tolerance distance is 250 m. Therefore, because of its high accuracy and number of destinations estimated, we choose 250 m as the tolerance distance for phases 21 and 22.

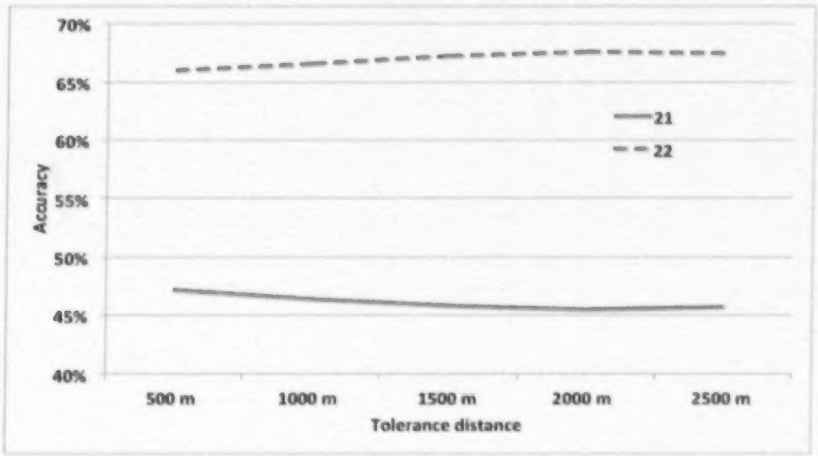

Figure 9 Calibration for tolerance distance of phases 21 and 22 (1) – estimated accuracy

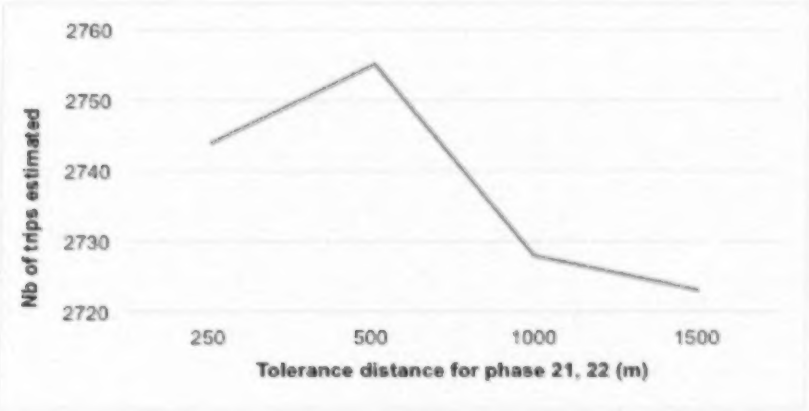

Figure 10 Calibration for tolerance distance of phases 21 and 22 (2) – total number of trips estimated

5.3 Results improvement

Figure 11 compares the accuracy before and after the calibration. For each phase, the accuracy increases, especially for phases 12 and 21. This could indicate that passengers get off at a stop near to their home, but the stop choice is looser for activity locations. Overall, the calibration helps to increase the accuracy by around 1.4 percentage points.
It is worth mentioning that the number of trips estimated is also slightly increased after calibration (+0.4%). This result is a combination of the decrease of the number of trips estimated by phases 11, 12 and 13, and the increase of the number of trips estimated by phases 21 and 22.

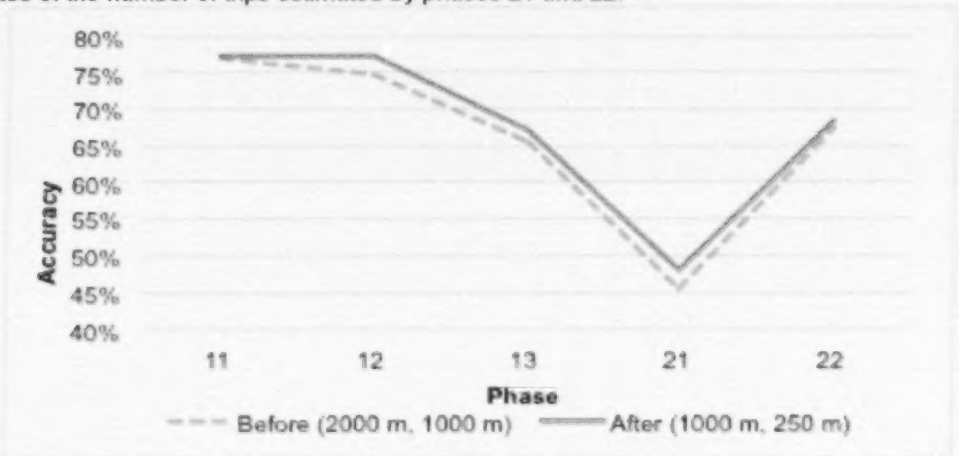

Figure 11 Comparison of accuracy before and after calibration

6 Conclusion

In this paper, we propose a method to calibrate a destination estimation algorithm, and we use a database taken from a smart card data set from Brisbane, Australia to validate this algorithm. The result shows that the accuracy is 65.76% and 79.17% for acceptable distances 0 and 400 m, respectively. We also proposed a method to calibrate the tolerance distance used in the algorithm. We chose the best tolerance distance as 1000 m for phases 11, 12 and 13 (the trip sequence part of the algorithm) and 250 m for phases 21 and 22 (unlinked trip part of the algorithm). After the calibration, the accuracy of estimation increased by 1.38% and the number of trips estimated increased by 0.4%.
However, for phase 21 (destination found with the improved method, where the unlinked trip has several potential destinations), the accuracy has been much improved with an acceptable distance increased to 400 m. This may lead to a conclusion that in further works, the algorithm should be calibrated at different tolerance distances for each of its phases, and additional data should be tested. It is also likely that calibration depends on the smart system site, because it may rely on the travel behaviour that is different from one city to another. At this time, there is no way to be assured that the calibrated parameters found for Brisbane would be the best for the case of Gatineau, where the algorithm was first developed.

Acknowledgements: The authors wish to acknowledge the supporters of this study, which are the Société de Transport de l'Outaouais, the Natural Science and Engineering Research Council of Canada (project RDCPJ 446107 12), Thalès Research and Technologies, and the Queensland Department of Transport and Main Roads.

References

Devillaine, F., Munizaga, M., & Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. Transportation Research Record: Journal of the Transportation Research Board, 2276(1), 48-55.

Gordon, J., H. Koutsopoulos, N. Wilson, and J. Attanucci (2013). Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transportation Research Record: Journal of the Transportation Research Board, 2343: 17-24.

He L & Trépanier M (2015). Estimating the destination of unlinked trips in public transportation smart card fare collection systems, Proceedings of the Annual Meeting of the Transportation Research Board, Washington, D.C.

Lee S & Ali A (2014). Analyzing subway origin-destination flows by utilizing smart transit card data: case of Seoul metropolitan area. 1st International Workshop on Utilizing Transit Smart Card Data for Service Planning, Gifu, Japan, July 2-3.

Morency C, Trépanier M, & Agard B (2007) Measuring transit use variability with smart-card data. Transport Policy 14(3):193–203.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies, 24, 9-18.

Nassir, N, Hickman, M, & Ma, Z (2015). Activity detection and transfer identification for public transit fare card data. Journal of Transportation, doi: 10.1007/s11116-015-9601-6.

Nassir, N., Khani, A., Lee, S.G., Noh, K., & Hickman, M. (2011). Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system, Transportation Research Record: Journal of the Transportation Research Board, 2263 (1), pp. 140-150.

Pelletier M.-P., Trépanier M., & C. Morency, Smart card data in public transit: A review, Transportation Research C: Emerging Technologies, 19(4), 557-568, 2011.

Shimamoto H (2014). Generation and calibration of transit hyperpath using smart card data. 1st International Workshop on Utilizing Transit Smart Card Data for Service Planning, Gifu, Japan, July 2-3.

Spurr, T., Chapleau, R., & Piché, D. (2014). Discovery and Partial Correction of Travel Survey Bias Using Subway Smart Card Transactions. In Transportation Research Board 93rd Annual Meeting (No. 14-4665).

Trépanier M, Morency C (2010) Assessing Transit Loyalty with Smart Card Data. 12th World Conference on Transport Research, Lisbon, Portugal.

Trépanier M, Morency C, Agard B (2009), Calculation of transit performance measures using smartcard data, Journal of Public Transportation, 12(1), 79-96.

Trépanier M, Morency C, Agard B, Descoimps É., Marcotte J.-S. (2012), Using smart card data to assess the impacts of weather on public transport user behavior, CASPT12 - Conference on Advanced Systems for Public Transport, Santiago, Chile, July 23-27.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems, 11(1), 1-14.